Liikennevirasto
Finnish Transport Agency

5 · 2014

RESEARCH REPORTS OF THE
FINNISH TRANSPORT AGENCY

HENRI SINTONEN

Analysis of the Predictability of Traffic during Congestion

Henri Sintonen

Analysis of the Predictability of Traffic during Congestion

Research reports of the
Finnish Transport Agency 5/2015

Finnish Transport Agency
Helsinki 2015

Photograph on the cover: Kari Hiltunen

Online publication (pdf) (www.liikennevirasto.fi)
ISSN-L 1798-6656
ISSN 1798-6664
ISBN 978-952-317-053-7

Finnish Transport Agency
P.O. Box 33
FIN-00521 HELSINKI, Finland
Tel. +358 (0)295 34 3000

Henri Sintonen: Analysis of the Predictability of Traffic during Congestion. Finnish Transport Agency, Traffic Services. Helsinki 2015. Research reports of the Finnish Transport Agency 5/2015. 32 pages. ISSN-L 1798-6656, ISSN 1798-6656, ISBN 978-952-317-053-7.

Keywords: traffic, congestion, travel time, predictability, analysis

Summary

Information about the speed of congestion alone may be insufficient to produce a high quality of service for road users, as some vehicles could pass the congested portion of the road network faster or slower than the mean speed would suggest. Because of this variation knowing the predictability of traffic is also important.

In this study the predictability of travel time was studied under different levels of congestion. Both recommended and new indicators of predictability were used. Moreover, a regression analysis was performed and indicator correlations were studied. Because the data consisted of anonymised observations from an automatic registration plate recognition system, only individual links were analysed. For longer distances, a method to combine data from consecutive links was developed. An effective and automated outlier removal method based on median was presented.

The results show that the predictability gets poorer with increased congestion, but exactly how much poorer varied depending on the indicator or method used and the road segment inspected. Good indicators were, in the end, scarce, so more research is needed. However, the best indicator turned out to be the median speed divided by the standard deviation (calculated from the median), but even this showed irregular behaviour, the authenticity of which is hard to evaluate. A threshold based assessment of congestion proved to be problematic, as it produced erroneous results. Joining the data of consecutive links seemed promising, but the nature of the available data set did not allow for a careful study of its efficacy.

Henri Sintonen: Analyysi liikenteen ennustettavuudesta ruuhka-aikoina. Liikennevirasto, liikenteen palvelut -osasto. Helsinki 2015. Liikenneviraston tutkimuksia ja selvityksiä 5/2015. 32 sivua. ISSN-L 1798-6656, ISSN 1798-6656, ISBN 978-952-317-053-7.

Avainsanat: liikenne, ruuhka, matka-aika, ennustettavuus, analyysi

Tiivistelmä

Korkean liikenteen palvelutason takaamiseksi ei välttämättä riitä tieto ruuhkaisuudesta. Osa ajoneuvoista voi ajaa ruuhkaisen tieosuuden läpi nopeammin tai hitaammin kuin ruuhkan keskinopeus antaisi olettaa. Tämän hajonnan takia matka-ajan ennustettavuus on myös tärkeä mittari.

Tutkimuksessa tutkittiin kuinka ennustettavaa liikenteen matka-aika on, kun ruuhkan taso vaihtelee. Tähän käytettiin suositeltuja ennustettavuuden indikaattoreita, sekä kokeiltiin uusia. Tutkimuksessa suoritettiin myös regressioanalyysi, sekä tutkittiin kuinka indikaattorit korreloivat keskenään. Koska tutkimuksessa käytetty data oli anonyymejä havaintoja automaattisesta rekisterikilven tunnistusjärjestelmästä, pystyttiin analyysi suorittamaan vain yksittäisillä linkeillä. Pidempien matkojen tutkimiseen kehitettiin menetelmä yhdistää peräkkäisten linkkien yksittäisistä havainnoista koostunutta dataa. Suuresti poikkeavien havaintojen poistamiseen esiteltiin toimiva ja automaattinen mediaaniin pohjautuva menetelmä.

Tuloksien mukaan liikenteen ennustettavuus laskee ruuhkan kasvaessa, mutta tarkalleen ottaen kuinka paljon laskua on, vaihtelee suuresti. Tämä riippui paljon mitä indikaattoria tai menetelmää käytettiin ja mitä tieosuutta tarkasteltiin. Hyvien indikaattorien määrä oli lopulta vähäinen, joten lisätutkimuksille on tarvetta. Kuitenkin näistä parhaaksi osoittautui mediaaninopeuden suhde keskihajontaan (mediaanista laskettuna), mutta tämänkin kohdalla havaittiin epäsäännöllistä käyttäytymistä, jonka todenperäisyyttä on vaikea arvioida. Ongelmia aiheutti myös raja-arvoihin perustuva ruuhkaisuuden määritelmä sen tuottamien virheiden takia. Peräkkäisten linkkien yhdistämismenetelmä vaikutti lupaavalta, mutta käytössä olevalla datalla ei perusteellista tutkimusta sen toimivuudesta kyetty tekemään.

Henri Sintonen: Analys av hur man kan förutsäga trafiken i rusningstider. Trafikverket, trafiktjänster. Helsingfors 2015. Trafikverkets undersökningar och utredningar 5/2015. 32 sidor. ISSN-L 1798-6656, ISSN 1798-6656, ISBN 978-952-317-053-7.

Sammanfattning

Vetskapen om trafikrusning räcker nödvändigtvis inte till för att garantera en god servicenivå i trafiken. En del av fordonen kan köra genom trafikrusningen snabbare eller långsammare än vad medelhastigheten i trafikrusningen skulle ge skäl att anta. På grund av denna spridning är också förutsägbarheten för restiden en viktig mätare.

I undersökningen tog man reda på hur bra restiden kan förutsägas när rusningstrafikens nivå varierar. För detta använde man rekommenderade indikatorer för förutsägbarheten samtidigt som man prövade nya. I undersökningen gjorde man också en regressionsanalys samt undersökte hur indikatorerna korrelerar sinsemellan. Eftersom den data som användes i undersökningen bestod av anonyma observationer från det automatiska systemet för identifiering av registerskyltar, kunde analysen bara göras för enskilda länkar. För att kunna undersöka längre resor utvecklade man en metod där man förenar data från enskilda observationer från på varandra följande länkar. För att eliminera observationer som avviker stort från varandra, presenterade man en fungerande och automatisk metod som baserar sig på medianen.

Enligt resultaten blir förutsägbarheten sämre ju svårare rusningstrafiken är, men exakt hur mycket sämre, varierar stort. Detta berodde i stor utsträckning på vilken indikator eller metod som användes och vilken vägsträcka som granskades. Det finns sist och slutligen få indikatorer som är bra och därför finns det behov av ytterligare undersökningar. Den bästa indikatorn visade sig vara medianhastighetens förhållande till standardavvikelsen (räknat från medianen), men också denna metod påvisade ett oregelbundet beteende, vars riktighet är svår att bedöma. Definitionen av rusningstrafik, som baserade sig på gränsvärden, orsakade också problem på grund av alla de fel som uppstod Metoden att förena på varandra följande länkar verkade lovande, men man kunde inte göra en grundlig undersökning om dess funktionalitet med den data som fanns att tillgå.

Foreword

In order to accurately evaluate the flow of traffic, knowing which factors influence the predictability of the time of arrival is crucial. This information is needed by industries that rely on accurate travel time predictions, by traffic management that monitors the road network and by any road user who wants to plan the most efficient routes. In this study, commissioned by the Finnish Transport Agency, data from real world observations were analysed to assess indicators for the predictability of traffic and see how they behave in and out of congestion.

The study was produced by Henri Sintonen from the VTT Technical Research Centre of Finland under the guidance of Kari Hiltunen and Dr. Risto Kulmala from the Finnish Transport Agency and Dr. Satu Innamaa from the VTT Technical Research Centre of Finland.

Helsinki, January 2015

Finnish Transport Agency
Traffic Services

Sisällysluettelo

1 Introduction

To ensure a good quality of service for road users the predictability of traffic is important. Knowing whether the journey will include congestion is crucial both when planning and during a trip, as the selected route could be modified during the trip based on new information. Just knowing how congested a section of the road network is might not be enough. Some vehicles might be able to pass the congested area quickly while for the others the trip might take a long time. This variation will have an impact on predictability. This study was designed to address the research question of how predictable travel time is between two points of the road network during congestion.

2 Methods

2.1 Data

2.1.1 Raw data

The study was based on travel time data collected in 2013 in the Greater Helsinki region. The data consisted of individual anonymised observations from an automatic number plate recognition system. Winter months were discarded to exclude the effects of weather; 3 months selected from other times of the year were May, August and October. Different links were used for different analyses and are described alongside the results in Section 3. Each observation consisted of the date and time of the observation and the measured speed and travel time of the vehicle. The observations were used to calculate the median speed of the traffic (Section 2.1.3) after outliers were removed from the data (Section 2.1.2).

2.1.2 Outlier Removal

The raw data included clear outliers that needed to be filtered out so as not to affect the results. Removal was done as follows: For each road segment r_i the data was divided into 30-minute time segments. First, the median absolute deviation (MAD) was calculated for the speed observations V_{d,r_i} of each time segment d:

$$MAD_{d,r_i} = 1.4826 * median(|v_i - median(V_{d,r_i})|), \text{ where } v_i \in V_{d,r_i}$$

The scale factor $K = 1.4826$ makes the MAD an estimator of the standard deviation. A median based measure was chosen, because procedures based on the mean (such as standard deviation threshold approaches) can be affected by extreme outliers (Leys et al., 2013). A modified z-score was calculated for each observation by dividing the differences between the speed value and the median speed of the observations by the MAD.

$$z_{i,r_i} = \frac{v_i - median(V_{d,r_i})}{MAD_{d,r_i}}$$

If the absolute value of this z-score exceeded the threshold of 3.5, the observation was marked as an outlier. In addition, if there were less than 10 observations in a segment, each observation with a speed below 30 km/h was also marked as an outlier. An example of the outcome of this procedure is shown in Figure 1. A few observations that could still be considered outliers were not removed, but given the amount of non-outlier data this is not problematic, and the last outliers would have had to be removed manually. As a systematic, fully automatized method was targeted the applied method was considered successful. Thus, the MAD based method was used to automatically remove outliers from the analyses.

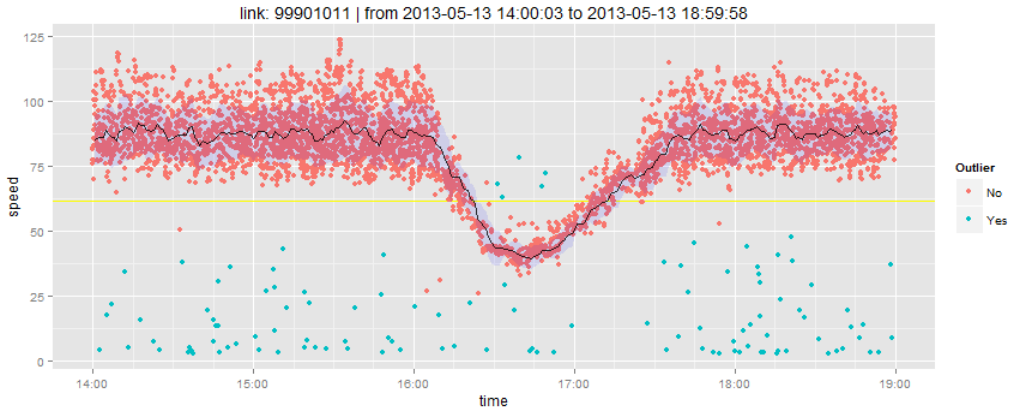

Figure 1 *Example of outlier removal. The black line is the 5 minute median; the blue shading on either side is the standard deviation (from the median) (see Section 2.1.3). The yellow line is the threshold for congestion.*

2.1.3 Data Preparation

Using the remaining observations after cleaning the data as described above, for each minute t and road segment r_i a median M_{t,r_i} of the (travel) speed (or travel time, used interchangeably) of the observations from the past 5 minutes was calculated. If there were fewer than eight observations in the 5-minute time window, the median speed was replaced with the free flow speed. The standard deviation (from the median M_{t,r_i}) σ_{t,r_i} was also calculated for each minute and road segment.

At the beginning of the study, the window was determined to be congested if the median travel time was at least 30% higher than in free flow, but this was changed following the results described in Section 3.1. The final congestion classification is as follows: The road section is considered congested during a 5 minute period if the median travel time was at least 50% higher than in free flow. Otherwise, the road section is considered to be in a state of transition or in light congestion if the median travel time was at least 15% higher than free flow travel time.

2.2 Indicators

Because the main research question on the predictability of travel time between two points of the road network during congestion cannot be directly answered using the available data set, a number of indicators were devised to infer the answer indirectly. Each indicator can be calculated for different time windows or a combination of them (e.g. only periods of congested traffic for 2 days).

2.2.1 Correlation between the Median and the Standard Deviation

In order to better understand the relationship between the median speed and the deviation of the observations from it, Pearson's correlation coefficient r was calculated for each of the pairs M_{t,r_i} and σ_{t,r_i}. The coefficient was also calculated separately for congested and non-congested traffic.

2.2.2 Median of the Standard Deviation

The standard deviation σ_{t,r_i} varies from minute to minute, so to compare the deviation during congestion with the deviation during non-congested traffic the median of the standard deviations σ_{t,r_i} was calculated for both instances.

2.2.3 Difference between Mean Speeds of Fast and Slow Vehicles

Fast vehicles were determined to be the fastest 15% and slow vehicles the slowest 15% of the observations of a time segment *d*. For both of these groups the mean speed was calculated. The final indicator was the difference between the mean speeds. The indicator can also be calculated separately for both congested and non-congested traffic. A large difference would indicate a high disparity in the speeds, which could make predictability worse.

$$Diff_{d,r_i} = mean\left(\{v | v \in V_{d,r_i}, v > V_{d,r_i}^{85\text{th}}\}\right) - mean\left(\{v | v \in V_{d,r_i}, v < V_{d,r_i}^{15\text{th}}\}\right),$$

where $V^{p\text{th}}$ refers to the *p*th percentile of V.

2.2.4 Median Absolute Deviation (MAD)

The median absolute deviation was calculated by taking the median of the absolute values of the differences between the observations v_j and the median of the medians of the speeds. That is,

$$MAD_{d,r_i} = median\left(\left|v_j - median\left(M_{t,r_i}\right)\right|\right), t \in d, v_j \in V_d$$

The indicator was also calculated separately for both congested and non-congested traffic. A high MAD value would indicate that the observations deviate greatly from the median, which could make predictability worse.

2.2.5 Indicators Recommended by the Finnish Transport Agency

Commissioned by the Finnish Transport Agency, Metsäranta et al. (2013) suggested using two indicators for the dispersion of travel times. The first is the standard deviation of the speed observations divided by the mean speed (SDM from here on). In this study the mean speed is replaced with the median speed:

$$SDM_{t,r_i} = \frac{\sigma_{t,r_i}}{M_{t,r_i}}$$

Since the data was prepared by calculating medians of 5-minute time windows, but the analysis will be done for larger time windows d, the indicator will be the average of these values:

$$SDM_{d,r_i} = mean(SDM_{t,r_i}), \text{ where } t \in d$$

The second indicator is the Planning Time Index (PTI), which is defined as the 0.95 percentile travel time divided by the free flow travel time. The interpretation is roughly that when the PTI is 1.5 a trip that normally takes 30 minutes takes 1.5*30 = 45 minutes.

Metsäranta et al. (2013) determined a classification for these indicators as shown in Table 1. The same color codes will be used in result tables to indicate predictability.

Table 1 *Classification of the predictability of travel time for SDM and PTI indicators (Metsäranta et al., 2013)*

Indicator	Excellent	Good	Satisfactory	Poor	Very poor
SDM	< 20%	≥ 20 % < 30 %	≥ 30 % < 40 %	≥ 40 % < 50 %	≥ 50 %
PTI	< 1.2	≥ 1.2 < 1.3	≥ 1.3 < 1.4	≥ 1.4 < 1.5	≥ 1.5

2.3 Link Combination Analysis

To further assess an answer to the research question, longer routes than the links in the travel time system should have been analysed therefore data from multiple consecutive links were to be combined. The anonymous nature of the observations posed difficulties in this regard and the aggregate statistics of consecutive links were combined instead. The idea was to simulate acquiring data from the same vehicles as they progress in traffic. This was conducted as follows:

A list of consecutive links was manually selected based on the number of observations and a starting time was selected from a period of congestion. For each link median travel time was calculated for a time period of 10 minutes starting from the starting time and using only the congested data points (minutes for which the medians from the past 5 minute were calculated) in order to know the speed of the vehicles in the congested traffic.

$$T = median(M_{t,r_i}), \text{ where } t \in [s, s + 10\text{min}], \text{ where } s \text{ is the starting time}$$

Next, the indicators were calculated for the link from a time period of $[s, s + T]$. For the next link in the list the starting time was set at $s + T$.

If for some link in the list there were not enough congested data points in the $[s, s + 10\text{min}]$ time period, the travel time was calculated from transition or light congestion data points. If that data was not available either, all data points of the time period were used to calculate the travel time. Lastly, if there were not enough vehicles at all during the time period, the free flow travel time for the link was used.

The reported results include the mean of the indicators of each link. The mean of the 5 minute medians and the mean of the standard deviations of each minute were also calculated so that they could be combined into the grand mean μ and grand standard deviation σ of the whole list of links. The data of the means does not overlap so the grand mean and standard deviation can be calculated in the standard manner:

$$\mu = \frac{\sum_i n_i \mu_i}{\sum_i n_i}$$

$$\sigma = \sqrt{\left(\frac{\sum_i n_i (\sigma_i^2 + \mu_i^2)}{\sum_i n_i}\right) - \mu^2}$$

where n_i is the number of 5 minute medians used, μ_i the mean of the 5 minute medians and σ_i the mean of the standard deviations of the minutes in the time period. Using these we can calculate the SDM indicator for the whole list of links with

$$SDM = \frac{\sigma}{\mu}$$

3 Results

Table 2 shows the links selected for analyses presented in Sections 3.1 and 3.2 as the road segments r_i due to their length and activity. Links used in other analyses are described in their own chapters.

Table 2 Links selected for the analyses in Sections 3.1 and 3.2

	Link ID	Name	Distance	Road	Free flow	
					Travel time, min:sec	Speed
1	312101	Klaukkala -> Herajoki	41608 m	3	21:48	115 km/h
2	312102	Herajoki -> Klaukkala	41608 m	3	21:48	115 km/h
3	99901032	Länsisalmi -> Muurala	32036 m	50	22:37	85 km/h
4	99901071	Veromies -> Riihikallio	9949 m	45	6:25	93 km/h
5	99901072	Riihikallio -> Veromies	9949 m	45	6:25	93 km/h
6	99901061	Käpylä -> Veromies	9321 m	45	6:17	89 km/h
7	99901011	Katajaharju -> Matinkylä	5790 m	51	4:21	80 km/h
8	99901012	Matinkylä -> Katajaharju	5790 m	51	4:21	80 km/h
9	99901041	Olari -> Kauniainen	5579 m	102	4:47	70 km/h
10	99901042	Kauniainen -> Olari	5579 m	102	4:47	70 km/h

3.1 Analysis of a Single Congestion Event

Analysis of a single congestion event was used to validate the method. The Olari-Kauniainen link on Main road 1 (99901041) from 2013-05-08 14:00:00 to 2013-05-08 18:00:00 had periods of both congestion and non-congestion. The event is summarised in Figure 2. For the non-congested minutes of the period with at least eight observations (during 14:00-15:30 and 17:15-18:00), the following results from the indicators were retrieved:

$M(M_{t,r_i})$	$M(\sigma_{t,r_i})$	SDM	PTI	Correlation	Diff	MAD
69.20	6.10	0.09	1.12	r(136) = 0.031 p=.72	9.18	17.95

The median amount of standard deviation from the median was 6.10 km/h and the difference between the mean speed of the fastest and slowest vehicles was 9.62 km/h. Together with the very low SDM and PTI the traffic could be considered predictable (green cells indicating excellent predictability) based on Table 1. The correlation test between the speed and the deviation was not statistically significant.

Next, the results were compared to the time window where the median speed falls under approximately 54 km/h, which was the threshold for congestion (15:30-17:15):

$M(M_{t,r_i})$	$M(\sigma_{t,r_i})$	SDM	PTI	Correlation	Diff	MAD
28.69	3.39	0.11	2.92	r(105) = 0.84 *	22.70	45.56

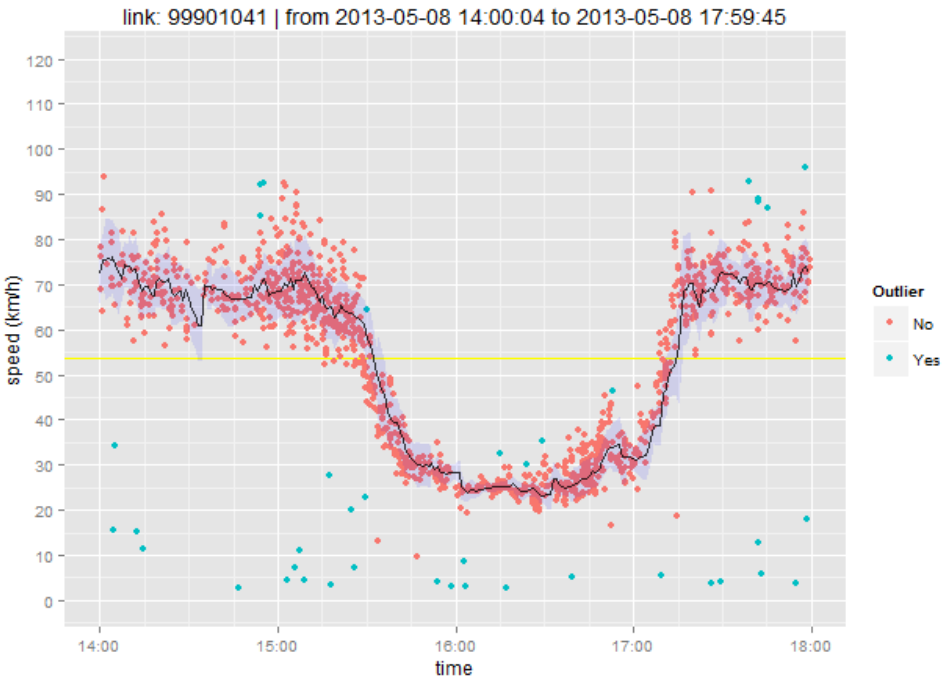

Figure 2 *Selected link and time period for the analysis. The black line is the 5 minute median speed; the blue shading on either side is the standard deviation (from the median) and the yellow line is the threshold for congestion.*

The standard deviation decreased while the PTI and the difference between slow and fast vehicles increased. SDM remained roughly the same (excellent, (Table 1)). The PTI rating moved from excellent to very poor. The Diff and the MAD increased dramatically, which indicates that throughout the congestion period there were multiple observations that deviated greatly from the median. There was strong and statistically significant correlation between the standard deviation and the median speed. However, it should be noted that the transition to congestion and then back to regular traffic are included in the congested traffic. Looking at the congested period between transitions (15:45-17:00), we get the following results:

$M(M_{t,r_i})$	$M(\sigma_{t,r_i})$	SDM	PTI	Correlation	Diff	MAD
26.60	2.40	0.10	2.96	r(75) = 0.73 *	8.39	3.83

The standard deviation continued to decrease slightly. As noted, both PTI and SDM are taken as indicators for predictability. Thus, from the results, the more than doubling of the PTI implies that congested traffic was much harder to predict (classification dropped from excellent to very poor), while the SDM remained excellent and did not change much during any of the periods examined in this section.

The difference between fast and slow vehicles dropped dramatically, as did the MAD. All the rest of the indicators produced similar results compared to the whole congestion period. Thus the Diff and MAD were misleading when calculated from traffic that was not non-congested if the desired interpretation is some sort of deviation from the median. They did, however, imply that during the time window there was a transition period or periods.

As big changes in speed can be expected during the transition from free flow to congestion, it was decided that the congested traffic would be split into two classes. For the following chapters, the threshold for congestion became 1.5 times the free flow travel time and the thresholds for transition or light congestion was 1.15 times the free flow travel time.

However, this was not a perfect solution to the problem either. The moment at which the travel time leaves the transition period and stagnates to congestion varies widely in regard to the link in question and the time of congestion. Selecting the threshold for multiple links and multiple days can be used as a compromise. The difficulty is clearly shown in Figure 3, where the peaks vary widely. Setting the threshold for congestion low will make it seem as if there was not much congestion along the road segment and setting it higher will make the median absolute deviation high.

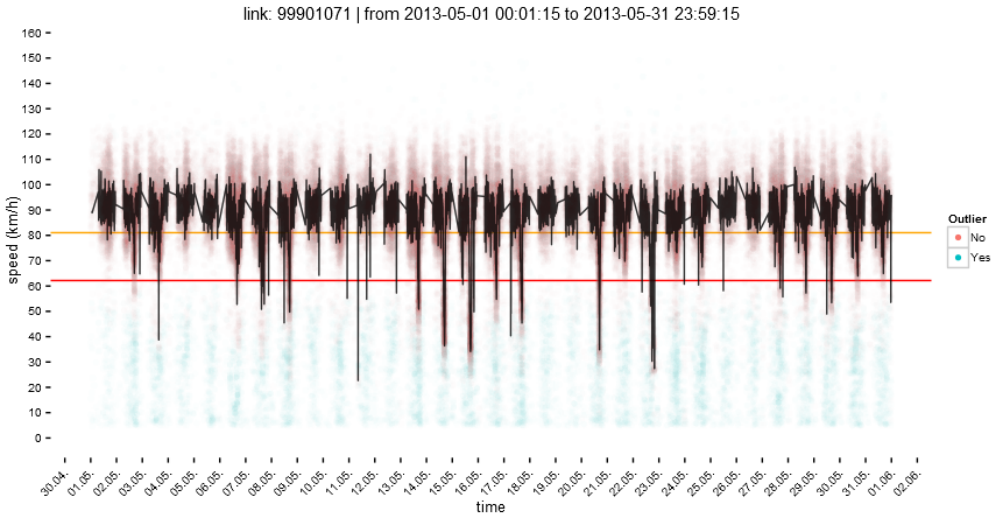

Figure 3 *Observations covering the month of May 2013 for link 99901071 with the new threshold for congestion. The observations are faded to show the median speed with more clarity. The orange line is the threshold for the transition or light congestion and the re*

For example, the new thresholds (about 61 km/h for transition or light congestion and 47 km/h for congestion) were not optimal for the congestion studied in this section. Unlike above when the manually selected congested period between transitions was examined, here the MAD remained high for the congestion:

Congested	M(M_{t,r_i})	M(σ_{t,r_i})	SDM	PTI	Correlation	Diff	MAD
No	69.21	6.10	0.09	1.12	r(136) = 0.03 p=0.72	9.18	17.95
T/L	51.63	7.03	0.14	1.50	r(11) = 0.26 p=0.39	9.72	28.30
Yes	28.45	3.26	0.11	2.94	r(95) = 0.81 p=0	14.39	45.92

However, these thresholds were deemed to be suitable, because they marked only a few clear transition peaks as just transition or light congestion (Figure 3).

3.2 Over a Month

To apply the enhanced method and get more general results, the analysis was run for all the data for one month on each link. The data was divided into three subsets. The subset for congested traffic consisted of all the minutes marked as congested during the data preparation process. The transition or light congestion subset was formed similarly. The last subset consisted of minutes that were not congested or in a transition period, but had more than eight observations. The indicators presented in Section 2.3 were calculated for both groups.

3.2.1 May 2013

The results for May 2013 are shown in Table 3. The data was available through May 25th.

Table 3 *Analysis results for the whole month of May 2013. The M() function is the median. In the correlation column the reporting follows the pattern of "r(degrees of freedom) = correlation coefficient *", where the degrees of freedom are always N-2 (N being the number of data points) and the asterisk is present if the results was statistically significant. "Diff" refers to the difference between fast and slow vehicles (Section 2.3.3). Values for the medians of the speed and the standard deviation, the difference indicator and for the MAD are in km/h. T/L refers to transition or light congestion.*

Link	Congested	M(M_{t,r_i})	M(σ_{t,r_i})	SDM	PTI	Correlation[1]	Diff	MAD
1	No	119.54	14.59	0.13	1.11	r(22194) = -0.15 *	16.01	16.38
	T/L	98.29	17.54	0.19	1.43	r(1881) = -0.12 *	16.45	30.81
	Yes	71.23	22.16	0.38	2.66	r(189) = -0.36 *	30.65	69.26
2	No	121.48	12.44	0.11	1.06	r(19140) = -0.30 *	11.00	9.21
	T/L	100.13	21.22	0.23	1.41	r(256) = -0.33 *	15.11	31.81
	Yes	49.80	7.40	0.22	3.82	r(63) = 0.56 *	43.64	106
3	No	84.12	11.86	0.16	1.13	r(5778) = -0.04 *	17.78	19.33
	T/L	66.86	14.22	0.23	1.48	r(1609) = 0.09 *	13.88	28.01
	Yes	43.36	13.47	0.34	2.84	r(2592) = 0.09 *	22.45	49.11
4	No	90.90	10.00	0.11	1.11	r(25130) = -0.12 *	11.18	12.42
	T/L	75.24	9.71	0.15	1.44	r(2120) = 0.15 *	13.09	22.20
	Yes	53.14	6.65	0.16	2.62	r(762) = 0.23 *	23.07	52.33
5	No	95.26	10.36	0.11	1.08	r(26157) = 0.01 p=0.07	13.01	11.08
	T/L	74.70	8.75	0.14	1.46	r(681) = 0.20 *	15.41	30.10
	Yes	49.50	9.28	0.43	5.97	r(391) = -0.18 *	38.06	66.71
6	No	92.44	10.15	0.11	1.04	r(35435) = 0.12 *	10.63	11.45
	T/L	67.11	7.30	0.11	1.47	r(746) = 0.29 *	14.08	37.18
	Yes	47.26	6.14	0.14	3.23	r(787) = 0.45 *	26.95	66.23
7	No	85.78	8.51	0.10	0.98	r(26844) = 0.22 *	8.92	8.77
	T/L	62.97	6.68	0.11	1.46	r(164) = -0.02 p=0.77	12.38	33.81
	Yes	33.76	3.89	0.51	22.28	r(175) = 0.06 p=0.44	47.70	77.13

[1] The following categorisation for the absolute values of the correlation coefficients is used here: > 0.8 Very strong, 0.5-0.8 Strong, 0.3-0.5 Moderate, 0.1-0.3 Modest, < 0.1 Weak

8	No	87.95	8.03	0.09	0.96	r(22268) = 0.38 *	12.79	8.73
	T/L	62.59	7.03	0.11	1.46	r(129) = 0.42 *	12.77	37.04
	Yes	40.16	3.99	0.18	3.44	r(181) = -0.49 *	21.86	70.30
9	No	71.22	4.95	0.07	1.06	r(17647) = -0.13 *	7.59	6.67
	T/L	53.60	5.63	0.11	1.48	r(752) = 0.21 *	11.66	25.39
	Yes	33.42	4.21	0.13	2.97	r(1602) = 0.32 *	18.83	54.21
10	No	72.90	8.27	0.11	1.03	r(30146) = 0.19 *	8.73	9.15
	T/L	56.98	7.81	0.15	1.46	r(284) = 0.09 p=0.12	10.79	23.41
	Yes	36.79	5.29	0.17	2.27	r(338) = 0.11 *	11.91	52.96

The results show that as the congestion progresses from no congestion to transition to congestion, the PTI increases. This is not surprising as the PTI was calculated by dividing travel time by free flow travel time. Thus as the travel time gets longer the PTI will increase. The PTI based predictability classification (Table 1) starts as excellent for all links when no congestion is observed. At the transition or light congestion level the predictability is poor for all links. During congestion the predictability is very poor for all links.

The SDM indicator (Table 3) also increases as congestion progresses, apart from links 2, 4 and 6, but there the difference is negligible. The SDM based predictability classifications always start as excellent for traffic with no congestion. For two links (2 and 3) the predictability is good during transition or light congestion and excellent for all remaining links. Finally, for congested traffic the predictability varies widely. For link 7 the predictability is very poor, for link 5 poor, for links 1 and 3 satisfactory, for link 2 good and for the remaining five links excellent. The results for the SDM indicator are shown in Figure 4; the distribution of SDM values for traffic without congestion are on the left, for transition or light congestion in the middle, and for congested traffic spread out on the right. Thus, during congestion the traffic is harder to predict and, because of the spreading, exactly how much harder the prediction is is also difficult to clarify. There is also considerable overlap in the 0.10-0.20 region of the SDM values, with some links being consistently harder to predict and some easier, even during congestion.

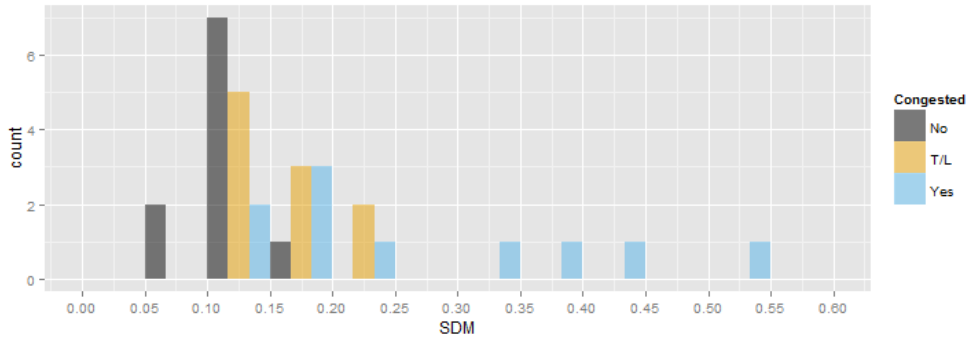

Figure 4 *Histogram of SDM indicator results for May 2013*

The median of the standard deviation (M(σ_{t,r_i}), Table 3) does not behave as neatly. It rises in link 1, drops in five of the 10 links and is mixed in four links as the congestion progresses.

The difference between fast and slow vehicles (Diff, Table 3) is mixed in links 3 and 8, but otherwise always increases. The values for no congestion vary from a minimum of 7.59 km/h to a maximum of 17.78 km/h with a mean of 11.76 km/h. The 1st and 3rd quartiles are 9.35 km/h and 12.96 km/h, respectively. For no transition or light congestion they vary from 10.79 km/h to 16.45 km/h with a mean of 13.56 km/h and the quartiles are 12.48 km/h and 14.85 km/h. All the values, apart from the maximum, increase, but only slightly. For the congestion the minimum is 11.91 km/h, maximum 47.70 km/h, mean 28.51 km/h and the quartiles 22.01 km/h and 36.21 km/h. The standard deviations of the difference for each congestion level are 3.23 km/h, 1.77 km/h and 11.43 km/h, respectively. The differences in absolute terms increase as the congestion progresses, but the deviation is smallest for the transition or light congestion period and clearly highest for the congestion period.

The MAD (Table 3) increases as the congestion increases in each link. The minimum, 1st quartile, mean, 2nd quartile and maximum for the no congestion level are 6.67 km/h, 8.87 km/h, 11.32 km/h, 12.18 km/h and 19.33 km/h, respectively. These values are close to those of the difference between fast and slow vehicles indicator above. For the transition or light congestion the values are 22.2 km/h, 26.05 km/h, 29.98 km/h, 33.31 km/h and 37.18 km/h. These are clearly higher than the values for the no congestion level and when compared to those of the Diff indicator. Finally, for the congested traffic the values are 49.11 km/h, 53.27 km/h, 66.42 km/h, 70.04 km/h and 105.99 km/h. The mean for the congested period is roughly twice that of the mean for the transition or light congestion. All the values are also much greater than those of the Diff indicator. The standard deviations of the MAD indicator for each congestion level are 3.87 km/h, 5.26 km/h and 16.77 km/h. The standard deviation of the no congestion traffic is close to that of the Diff indicator, but all the rest are higher. The narrowness of the transition or light congestion period is not seen in the MAD indicator as it was in the Diff indicator. The high values of the MAD might be due to the inadequacy of the threshold based congestion classification, as discussed in the previous section. If the congestion classifications were optimal and the MAD still high for the congestion period, it would mean that many observations deviated strongly from the median, which could be taken as a sign of difficulty predicting traffic. However, the former explanation seems more plausible.

The correlation between the median of the median speed and the median of the standard deviation behaves erratically. It varies from no correlation to strong correlation from link to link and congestion level to congestion level. Within one link it can show both a negative correlation and a positive correlation for different congestion levels in unpredictable ways.

3.2.2 August 2013

The August 2013 results paint a similar picture (Table 4). The median of the standard deviation and the correlation do not aid in understanding the congestion or its predictability. The SDM has seven links with excellent predictability in transition or light congestion traffic and three with good. This is one fewer excellent marks than in May 2013. Again the SDM varies greatly for congestion: four excellent, one good, two satisfactory, two poor and two very poor. Thus predictability is slightly worse than in May 2013. The histogram of the SDM values (Figure 5 Histogram of SDM values for August 2013) is structured the same way as before.

Table 4 Results for August 2013

Link	Congested	M(M_{t,r_i})	M(σ_{t,r_i})	SDM	PTI	Correlation	Diff	MAD
1	No	119.59	14.35	0.13	0.97	r(30035) = -0.25 *	18.67	14.97
	T/L	82.86	19.77	0.26	1.47	r(319) = -0.20 *	14.79	53.12
	Yes	57.41	26.92	0.50	2.41	r(98) = -0.31 *	19.45	90.71
2	No	121.48	12.30	0.11	0.88	r(26845) = -0.33 *	11.12	9.21
	T/L	80.62	17.00	0.23	1.46	r(62) = 0.03 p=0.83	14.15	60.44
	Yes	64.11	23.66	0.39	1.60	r(9) = -0.15 p=0.65	3.13	84.92
3	No	83.38	11.59	0.16	1.13	r(5618) = 0.01 p=0.39	17.02	18.87
	T/L	67.17	13.63	0.22	1.47	r(1742) = 0.04 p=0.07	13.86	26.61
	Yes	43.00	12.60	0.32	2.95	r(2455) = 0.12 *	22.66	48.81
4	No	91.37	9.48	0.11	1.10	r(25230) = -0.07 *	10.90	11.39
	T/L	76.21	9.55	0.14	1.45	r(1113) = 0.18 *	14.19	22.05
	Yes	56.40	7.28	0.15	2.00	r(354) = -0.07 p=0.21	12.33	49.46
5	No	95.00	10.80	0.12	1.07	r(30652) = 0.07 *	13.17	11.34
	T/L	76.21	11.76	0.17	1.44	r(430) = 0.05 p=0.33	12.48	27.87
	Yes	48.08	17.90	0.43	3.64	r(258) = -0.10 p=0.11	31.92	68.83
6	No	92.70	10.37	0.11	1.03	r(36728) = 0.07 *	10.31	11.48
	T/L	69.04	7.98	0.12	1.47	r(494) = 0.27 *	13.44	34.69
	Yes	52.84	6.74	0.13	2.24	r(191) = 0.35 *	16.08	58.71
7	No	86.13	8.45	0.10	0.98	r(30061) = 0.32 *	8.99	8.84
	T/L	65.55	8.27	0.13	1.45	r(160) = 0.25 *	11.06	31.05
	Yes	28.36	4.49	0.88	17.87	r(57) = -0.15 p=0.26	43.67	86.18
8	No	83.88	6.74	0.08	1.01	r(30970) = 0.18 *	8.38	7.14
	T/L	60.68	5.46	0.09	1.47	r(235) = 0.15 *	13.70	34.14
	Yes	44.07	4.40	0.13	5.30	r(245) = 0.29 *	27.23	58.77
9	No	75.08	5.07	0.07	1.00	r(17496) = -0.15 *	7.49	6.38
	T/L	54.58	6.12	0.12	1.47	r(339) = 0.08 p=0.13	11.52	29.78
	Yes	31.70	3.72	0.14	3.92	r(936) = 0.38 *	22.78	62.87
10	No	76.23	8.56	0.11	0.98	r(31287) = 0.19 *	8.32	8.90
	T/L	53.99	9.68	0.19	1.46	r(75) = 0.33 *	10.72	32.75
	Yes	36.92	8.02	0.27	2.34	r(99) = -0.18 p=0.08	14.62	58.05

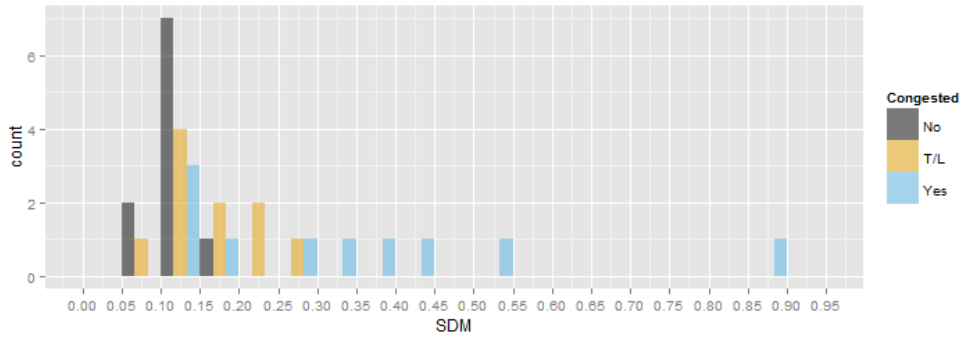

Figure 5 Histogram of SDM values for August 2013.

The PTI (Table 4) is poor for every link for transition or light congestion and very poor for every link during congestion. This is similar to May 2013. The Diff (Table 4) did not change much from May 2013, except for the mean, 1st quartile and 3rd quartile during congestion, which dropped by 7.12 km/h, 7.05 km/h and 10.09 km/h, respectively. Otherwise the changes were less than 2 km/h compared to May 2013.

The MAD (Table 4) increased for the transition or light congestion level by 2.30 km/h, 5.27 km/h, 1.24 km/h and 23.26 km/h for the 1st quartile, mean, 3rd quartile and the maximum. The minimum did not change much. During congested periods the mean and the minimum did not change much, but the change in the 1st quartile, 3rd quartile and the maximum was 4.95 km/h, 10.86 km/h and -15.28 km/h, respectively.

3.2.3 October 2013

Finally, Table 5 shows the results for October 2013. The SDM indicated that six links had excellent and four had good predictability for transition or light congestion periods. Thus predictability is slightly worse than in August or May 2013. The congested traffic varied again, with three excellent, three good, two satisfactory and one very poor. Link 10 did not have enough data for the congested period for the analysis. The histogram of SDM values (Figure 6) again shows a familiar structure, but the predictability of transition or light congestion is more spread out. The PTI (Table 5) gave poor predictability for transition or light congestion and very poor predictability for congested traffic, which is similar to previous months. Diff (Table 5) acted similarly as it did in May 2013, with changes of less than 3 km/h except for the maximum. The MAD (Table 5) also acted similarly to May 2013.

Table 5 Results for October 2013

	Congested	M(M_{t,r_i})	M(σ_{t,r_i})	SDM	PTI	Correlation	Diff	MAD
1	No	118.32	13.44	0.12	0.98	r(23070) = 0 p=0.95	21.61	15.37
	T/L	82.26	14.81	0.21	1.40	r(199) = -0.15 *	12.86	52.09
	Yes	57.92	28.75	0.51	3.26	r(53) = -0.20 p=0.15	24.62	88.16
2	No	119.64	12.14	0.11	0.96	r(26580) = -0.11 *	18.94	12.11
	T/L	79.72	18.98	0.25	1.48	r(129) = 0.10 p=0.27	17.16	58.33
	Yes	50.69	6.30	0.22	3.43	r(240) = 0.25 *	30.6	101.23
3	No	82.73	11.90	0.16	1.13	r(4768) = -0.03 p=0.05	16.56	20.91
	T/L	66.36	13.78	0.22	1.48	r(1800) = 0.10 *	13.93	27.05
	Yes	42.94	12.79	0.33	2.83	r(2718) = 0.06 *	22.3	45.04
4	No	90.33	9.26	0.11	1.11	r(20867) = -0.02 *	11.02	11.16
	T/L	77.19	9.02	0.13	1.45	r(1356) = 0.09 *	13.23	19.31
	Yes	45.51	5.88	0.23	9.49	r(298) = -0.06 p=0.3	33.17	63.97
5	No	93.64	10.73	0.12	1.10	r(29052) = 0.06 *	14.19	12.4
	T/L	74.96	8.99	0.15	1.47	r(1412) = 0.19 *	14.91	26.79
	Yes	49.33	6.24	0.32	5.12	r(872) = -0.24 *	34.63	63.02
6	No	92.82	10.52	0.11	1.04	r(34790) = 0.04 *	10.77	12.93
	T/L	68.00	8.57	0.13	1.47	r(999) = 0.24 *	14.51	35.12
	Yes	52.43	7.79	0.20	2.16	r(585) = -0.32 *	15.55	57.82
7	No	85.61	7.67	0.09	0.99	r(28400) = 0.13 *	8.72	8.61
	T/L	64.53	7.35	0.11	1.46	r(352) = 0.23 *	11.37	30.46
	Yes	48.59	5.37	0.12	8.17	r(123) = 0.75 *	32.33	54.1
8	No	82.91	7.12	0.09	1.02	r(28951) = 0.23 *	8.65	7.60
	T/L	62.97	5.59	0.09	1.46	r(375) = -0.06 p=0.28	12.35	29.27
	Yes	37.09	3.27	0.11	3.57	r(519) = 0.30 *	25.08	67.15
9	No	74.53	5.84	0.08	1.01	r(16324) = 0.08 *	8.55	6.74
	T/L	54.14	5.23	0.10	1.48	r(468) = 0.23 *	11.97	29.61
	Yes	34.13	3.38	0.11	3.24	r(801) = 0.46 *	20.58	58.88
10	No	75.67	8.47	0.11	0.99	r(26570) = 0.18 *	9.36	8.79
	T/L	58.31	10.11	0.24	1.49	r(35) = -0.76 *	12.43	25.5
	Yes	-	-	-	-	-	-	-

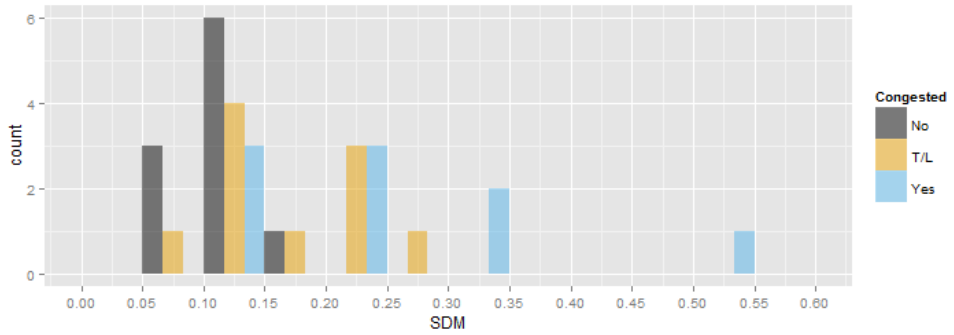

Figure 6 *Histogram of SDM values for October 2013.*

3.2.4 Summary of the Results

A summary of the M(σ_{t,r_i}), Diff and MAD results is presented in Table 6 and Figure 7. The quartiles, median and mean of the M(σ_{t,r_i}) and MAD increased and the Diff decreased during August. The other values were relatively stable from month to month.

Table 6 *Summary of medians of the standard deviations, Diff and MAD values of the analyses. All values are in km/h. Q = quartile.*

	Month	Min.	1st Q	Median	Mean	3rd Q	Max.
M(σ_{t,r_i})	May	3.89	6.66	8.39	9.58	11.49	22.16
	August	3.72	6.88	9.52	10.76	12.53	26.92
	October	3.27	6.24	8.57	9.63	11.90	28.75
Diff	May	7.59	11.72	13.98	17.95	21.10	47.70
	August	3.13	10.94	13.57	15.27	16.79	43.67
	October	8.55	11.97	14.51	17.31	21.61	34.63
MAD	May	6.67	13.41	30.45	35.91	52.80	106.00
	August	6.38	12.35	31.90	37.61	58.55	90.71
	October	6.74	12.93	29.27	36.19	57.82	101.23

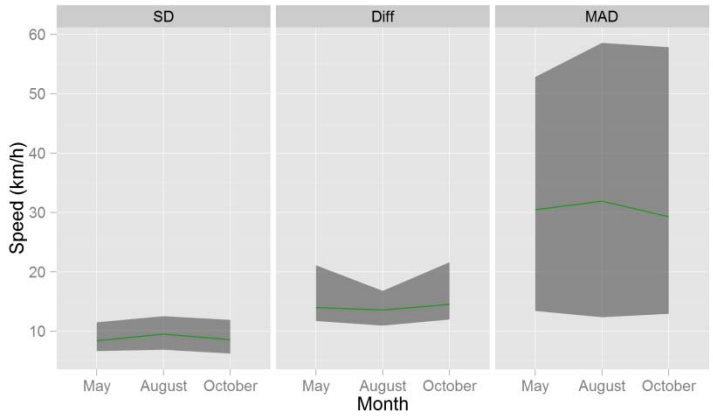

Figure 7 *Summary of the median and the 1st and 3rd quartiles of the analyses.SD is the median of the standard deviations M(σ_{t,r_i}). Data from Table 6.*

The SDM values are summarised in Table 7. During no congestion the predictability of each link was excellent (see Table 1). Links 1-3 were consistently characterized as harder to predict during transition or light congestion, but still had good predictability. Links 1, 3, 5 and 7 clearly had even poorer predictability for congested traffic, but the predictability varied, as also seen in Figure 8.

Table 7 *Summary of SDM values of the analyses. "No", "T/L" and "Yes" refer to congestion.*

Link	Month	No	T/L	Yes
1	May	0.13	0.19	0.38
	August	0.13	0.26	0.50
	October	0.12	0.21	0.51
2	May	0.11	0.23	0.22
	August	0.11	0.23	0.39
	October	0.11	0.25	0.22
3	May	0.16	0.23	0.34
	August	0.16	0.22	0.32
	October	0.16	0.22	0.33
4	May	0.11	0.15	0.16
	August	0.11	0.14	0.15
	October	0.11	0.13	0.23
5	May	0.11	0.14	0.43
	August	0.12	0.17	0.43
	October	0.12	0.15	0.32
6	May	0.11	0.11	0.14
	August	0.11	0.12	0.13
	October	0.11	0.13	0.20
7	May	0.10	0.11	0.51
	August	0.10	0.13	0.88
	October	0.09	0.11	0.12
8	May	0.09	0.11	0.18
	August	0.08	0.09	0.13
	October	0.09	0.09	0.11
9	May	0.07	0.11	0.13
	August	0.07	0.12	0.14
	October	0.08	0.10	0.11
10	May	0.11	0.15	0.17
	August	0.11	0.19	0.27
	October	0.11	0.24	-

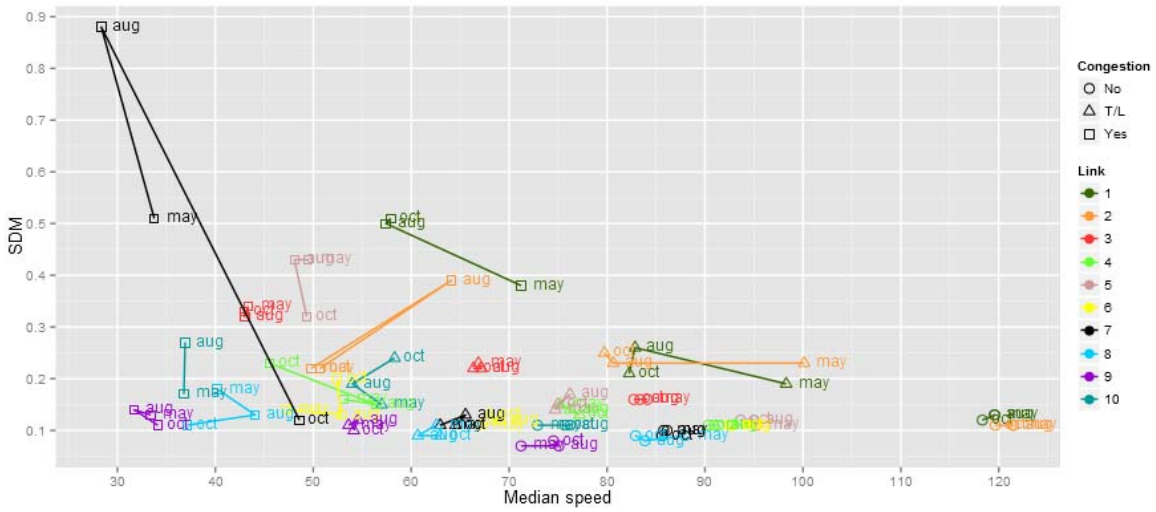

Figure 8 *Summary of SDM values together with median speed and congestion level for each month.*

3.3 Regression Analysis

A regression analysis was done for link 3 due to its consistent behaviour with regards to predictability (see Table 6). Analysing data from the whole month of May 2013 indicated that the speed of observations V_{r_i} was significantly predicted by the median speed M_{t,r_i} with the formula $V_{r_i} = 25.43 + 0.61 * M_{t,r_i}, R^2 = 0.21$. This line is visible in Figure 9. Splitting the month into days and calculating the same formula for each day and taking the mean of the results produced the formula $V_{r_i} = 28.11 + 0.57 * M_{t,r_i}, R^2 = 0.18$, which is close to the formula calculated from the data for the whole month.

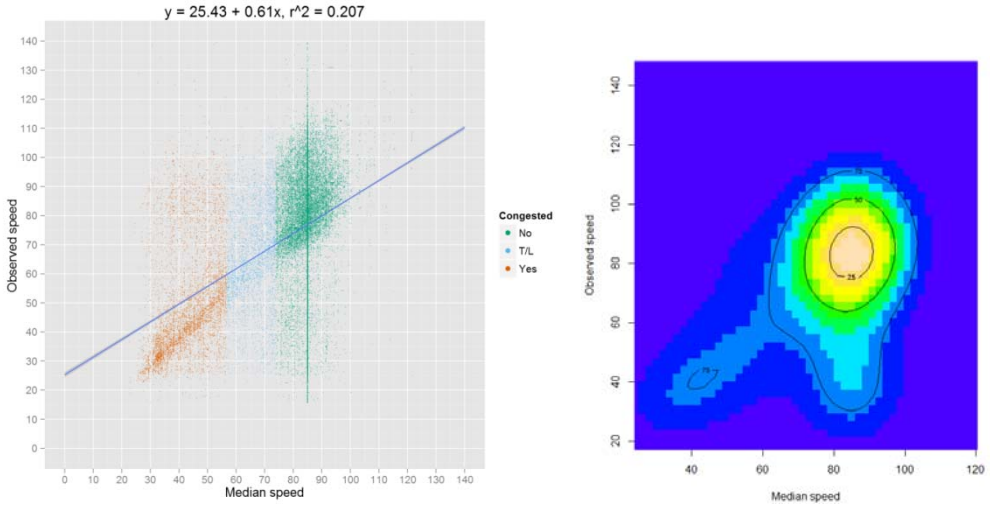

Figure 9 *Regression analysis of link 3, May 2013. Left: observations and regression line. The vertical line at 85 km/h is the free flow speed. Right: density plot of observations. The contour lines give the percentage (25%, 50% and 75%) of the number of observations present within each area.*

Figure 9 shows that the observations form a denser line-like shape that increases with the median speed. However, the fitted regression line is slightly off from the denser area. The y-intercept should be closer to zero. Removing over 14 200 data points with a median speed equal to the free flow speed causes only a slight change in the formula: $V_{r_i} = 20.46 + 0.72 * M_{t,r_i}, R^2 = 0.37$, thus the reason for the difference is probably the spreading of observations around the regression line, especially for those of no congestion. If 10% of the slowest and 10% of the fastest observations of each integer of the median speed are further removed, the y-intercept continues to decrease and the coefficient of determination greatly increases: $V_{r_i} = 14.84 + 0.80 * M_{t,r_i}, R^2 = 0.65$. The regression line seems to rotate around the dense point based on the spreading of the observations.

The spreading was studied using standard deviation (SD) and interquartile range (IQR, defined as Q3-Q1) after removing the data points with a median speed equal to the free flow speed, but those in the 10% of the fastest or slowest vehicles. The results are presented in Table 8. Both IQR and SD indicate that the spreading increases as congestion progresses. Thus regression type prediction is more difficult during congestion.

Table 8 *Spreading of observations from May 2013.*

Congested	IQR (km/h)	SD (km/h)
No	18.36	17.33
T/L	24.94	18.06
Yes	25.97	20.35

3.4 Indicator Correlations

The Diff and MAD indicators showed some similar behaviour in the previous sections, whereas the median of the standard deviation behaved erratically. To further study the co-behaviour of the indicators, the correlation between them was continued from the correlation between the median of medians and the median of the standard deviation already utilized in the previous sections.

Seventy-five percent of the most active links from an area roughly corresponding to the Greater Helsinki region were used for the analysis (151 links). For each of them the observations from May 2013 were divided into the three congestion classes and the indicators were calculated for each pair of link and congestion class. Pearson's correlation coefficient matrix was calculated for the indicators and the results are summarised in Table 9.

Table 9 *All correlations are statistically significant ($p < 0.05$) except those in red. Sample size was 289 (106 rows with no congestion, 96 with transition or light congestion and 87 with congestion) after removing cases with too few observations.*

	M(M_{t,r_i})	M(σ_{t,r_i})	SDM	PTI	Diff	MAD
M(M_{t,r_i})	1	0.40	-0.34	-0.45	-0.46	-0.62
M(σ_{t,r_i})		1	0.04	-0.19	-0.06	-0.23
SDM			1	0.60	0.25	0.35
PTI				1	0.46	0.42
Diff					1	0.76
MAD						1

As seen in previous sections the Diff and MAD seem to be strongly correlated, whereas the median of the standard deviation does not correlate well with anything. Apart from the median of the standard deviation all the indicators have a negative correlation with the median of medians meaning that when the speed of the traffic decreases the indicators increase. The Diff is as moderately correlated with the speed as is PTI, whereas the MAD alone is strongly correlated with the speed. PTI and SDM are well correlated with each other. The Diff and MAD are not as well correlated with the PTI or SDM, but are not far from it either.

Next the same procedure was done, but this time the congestion levels were calculated separately for each day of the month for each link. This increases the sample size of the correlation calculation, but each data point is calculated from a smaller sample of observations than before. The results are presented in Table 10.

Table 10 *Sample size 4790 (2685 no congestion, 1246 T/L and 859 congestion)*

	M(M_{t,r_i})	M(σ_{t,r_i})	SDM	PTI	Diff	MAD
M(M_{t,r_i})	1	0.27	-0.46	-0.40	-0.26	-0.47
M(σ_{t,r_i})		1	0.34	-0.18	0.20	-0.03
SDM			1	0.45	0.45	0.29
PTI				1	0.27	0.21
Diff					1	0.58
MAD						1

The most notable changes were: The SDM and median of the standard deviation were now moderately correlated, the PTI and SDM correlation dropped to moderate, the Diff and SDM correlation increased to moderate, and the Diff and PTI correlation dropped to modest.

3.5 Link Combination Analysis

A link combination analysis was performed as detailed in Section 2.3. The lists of links were selected based on their number of observations during congestion (or during transition or light congestion or during no congestion, whichever was the case), with the majority of 5 minute medians being classified as such. The list of links did not end up overlapping with the links selected for the previous analyses, unless otherwise stated below. The indicators, however, are calculated from all 5 minute medians. The initial starting times were selected such that the results of the first link on the list correspond well to the level of congestion chosen.

Munkkiniemi -> Kehä 1 -> Stensintie (Road 1)

Table 11 *Results of link combination analysis for the links Munkkiniemi to Kehä 1 and Kehä 1 to Stensintie. m() refers to using the mean. The last value in the table, SDM, is calculated from the first two (grand mean and grand standard deviation). Starting times are listed in Table 12. The number of observations are raw data points used only to calculate the 5 minute medians, which are then used in the link combination analysis.*

Congested	Grand mean	Grand sd	Observations	m(PTI)	m(SDM)	m(Diff)	m(MAD)	SDM
Yes	24.62	14.78	65	4.85	0.32	12.17	9.99	0.60
No	95.46	12.48	33	1.04	0.09	3.17	8.61	0.13

Table 12 *Initial starting times and following starting times calculated as detailed in Section 2.3.*

Congested	Initial starting time	Starting times of the following links	Ending time
Yes	2013-05-08 17:23:00	17:30:39	17:44:30
No	2013-05-08 15:15:00	15:17:17	15:20:11

The results of the first link combination are shown in Table 11 and Table 12. A specific moment at which the traffic was mostly in transition or light congestion in the combination was not found. The mean speed behaves as expected with lower speeds during congestion, as does the total travel time in Table 12. Thus the PTI indicates a jump from excellent to very poor predictability (see Table 1). The mean of the SDM

values of the minutes analysed, m(SDM), indicates a jump from Excellent predictability to Satisfactory predictability. The jump in the mean of the Diff values of the minutes analysed is almost fourfold during congestion compared with no congestion. The mean of the MAD values increased slightly during congestion, but in absolute terms remained quite low, as did MAD during the single event analysis in Section 3.1 and unlike in the analysis of whole months in Section 3.2.

The SDM, which was calculated from the grand mean and grand standard deviation for the whole month, jumped from excellent predictability during no congestion to very poor predictability during congestion. In numerical terms the SDM was almost twice that of the m(SDM) during congestion, but very similar during no congestion.

Pitäjänmäki -> Konala -> Varisto (Road 120)

The results of the second combination are presented below.

Congested	Grand mean	Grand sd	Observations	m(PTI)	m(SDM)	m(Diff)	m(MAD)	SDM
Yes	19.12	7.44	122	7.38	0.61	43.41	2.57	0.39
T/L	48.51	13.81	21	1.58	0.31	11.74	14.12	0.28
No	60.26	10.09	22	1.14	0.17	5.87	13.15	0.17

Congested	Initial starting time	Starting times of the following links	Ending time
Yes	2013-05-02 16:45:00	16:46:58	17:11:23
T/L	2013-05-02 17:15:00	17:16:58	17:22:12
No	2013-05-02 21:36:00	21:38:08	21:42:53

Here, during no congestion and transition or light congestion, the m(SDM) and SDM are close to each other. Both show excellent predictability for no congestion and are close to the transition point between good and satisfactory for transition or light congestion, albeit on different sides. During congestion m(SDM) is over 1.5 times the SDM. For congested traffic the SDM indicates satisfactory and m(SDM) very poor predictability. The PTI is excellent for no congestion and very poor for the other congestion levels.
The deviation based indicators are varied. The grand standard deviation and m(MAD) are highest for the transition or light congestion level, but m(Diff) is highest for the congestion level.

Käpylä -> Pakila -> Tammisto -> Veromies -> Riihikallio (Road 45)

Here the last link, from Veromies to Riihikallio, corresponds to link 4 in the previous analyses. The results for the third combination are presented below.

Congested	Grand mean	Grand sd	Observations	m(PTI)	m(SDM)	m(Diff)	m(MAD)	SDM
Yes	50.71	9.76	281	1.72	0.11	2.94	5.40	0.19
T/L	72.10	14.72	237	1.36	0.12	2.58	6.60	0.20
No	87.34	7.08	100	1.11	0.11	5.28	11.41	0.08

Congested	Initial starting time	Starting times of the following links			Ending time
Yes	2013-05-13 16:27:00	16:33:52	16:37:20	16:38:31	16:50:01
T/L	2013-05-13 17:20:00	17:24:24	17:27:46	17:28:58	17:36:28
No	2013-05-13 17:45:00	17:47:59	17:49:58	17:50:58	15:57:46

The m(SDM) does not change much from congestion level to congestion level (excellent). The SDM, however, is similar for congestion and transition or light congestion (close to the transition point between excellent and good), but much lower for no congestion (excellent). Based on previous results from this study the behaviour of SDM is more familiar, as the predictability is usually better for no congestion. It should be noted that the free flow speeds of the consecutive links were varied (80 km/h, 96 km/h, 100 km/h and 93 km/h, respectively).

Riihikallio -> Ruotsinkylä -> Ilola -> Veromies -> Käpylä (Road 45)

Finally, the results for the last link combination are presented below.

Congested	Grand mean	Grand sd	Observations	m(PTI)	m(SDM)	m(Diff)	m(MAD)	SDM
Yes	35.75	13.19	234	2.37	0.11	3.19	6.38	0.37
T/L	63.57	8.48	105	1.47	0.09	3.11	7.91	0.13
No	93.87	12.72	47	0.96	0.06	1.68	30.33	0.14

Congested	Initial starting time	Starting times of the following links			Ending time
Yes	2013-05-23 08:10:00	08:11:42	08:15:29	08:21:31	08:41:02
T/L	2013-05-23 07:00:00	07:02:06	07:04:30	07:06:41	07:12:58
No	2013-05-23 09:30:00	09:31:45	09:33:57	09:36:00	09:42:21

Here the standard deviations show a narrower transition period as seen above. The m(SDM) only shows slight changes, while the SDM has a much higher value for the congested period. As in the previous combination, the free flow speeds of the links varied (81 km/h, 100 km/h, 100 km/h and 89 km/h, respectively.

4 Conclusions and Discussion

This study was designed to assess whether the predictability of traffic between two points of the road network is poorer during congestion. Two indicators were provided: the Planning Time Index (PTI) and the standard deviation divided by the mean (median here) speed (called SDM here). The available data consisted of anonymous observations taken from short links between automatic number plate recognition measuring devices. Thus, observations between two points farther apart than the relevant measuring devices were not available, and the research question could not be answered straightforwardly.

To remedy this, longer active links were selected for the analysis and more indicators were devised to increase confidence in the results. These indicators were the difference between fast and slow vehicles, median absolute deviation and the correlation between median speed and the standard deviation. Moreover, a method was devised to connect the aggregate values of successive links to try to imitate the data needed to answer the research question.

Outliers from the data were removed using a median absolute deviation (MAD) based method, which turned out to be a good way of removing nearly all the outliers automatically. Two choices were made when applying it. First was the use of 30-minute non-overlapping time windows and the second was the threshold parameter that, in the end, determined which observation was an outlier and which was not. By tweaking these choices the method might be further improved.

First the observations were divided into two categories, congested and non-congested. This, however, was found to lead to major deviations in the speeds of the congested observations, as it included the transition to and from the congestion. Thus, a third category was created for the transition periods. This still turned out to be inadequate, as the classification of observations into the congestion categories was based on thresholds. Some congestion waves with transitions and stagnation at the peak congestion were labelled as transitions, because the median speed of the vehicles did not exceed the arbitrary threshold value. Moreover, some of the congestion peaks were much higher than the threshold value, which means that not only are the transition periods contaminated with stagnant periods, but also the congestion periods are contaminated with transition periods.

Because of this contamination some indicators might not be truthful. At least the median absolute deviation and the difference between slow and fast vehicles could be amplified. A high value for both these indicators would imply that there are many vehicles with vastly different speeds, but using the threshold based congestion classification it is unclear which of these two causes any given increase in these two indicators.

A more advanced congestion classifier should be used in future research. The improved method should detect congestion from the shape and evolution of the median speed. Then, afterwards, a threshold method could be used to only pick the shapes where the peak crosses the selected threshold. In this way even transitions that do not lead to a peak that exceeds the threshold could be removed from the set of transition periods.

Looking at how the indicators correlate with each other, the two recommended indicators had a moderate or strong correlation. The other indicators had either a slightly poorer correlation with the PTI and SDM or very similar. With proper congestion classification it might be possible to craft them into good indicators for predictability. For example, the MAD indicator might be used as a term in a ratio similar to standard deviation in SDM.

The results of the predictability analysis are very clear if we look at the PTI. Predictability is clearly poorer when the traffic is in a state of transition (or light congestion) than when there is no congested traffic. It is even worse (very poor) when looking at the congested traffic. However, the PTI might not be a good candidate for predictability indicator, as it only takes into account the increase in travel time.

The SDM gave similar results, but less clear and less harsh. It too implied that predictability worsens as congestion progresses, but the predictability was more spread out from excellent to very poor compared to the PTI. The difference between fast and slow vehicles and the median absolute deviation too implied that predictability gets worse during congestion, but, as said, it is unclear whether they actually show contamination of the congestion classification. The correlation between the median speed and standard deviation of the speeds turned out to be an inconsistent indicator.
In conclusion, the SDM seemed to be the best indicator in this study. The PTI can be challenged based on its definition and the correlation between the median and the standard deviation acted erratically. Difference between fast and slow vehicles and the MAD are promising, but not when used alone. They could be used as terms in some other, more complicated indicator. While the SDM was deemed best, it is hard to know how truthful it is, as objective ground truth regarding predictability was not available.

The results were compared from May 2013 to August 2013 to October 2013. From the predictability categories there was slight variation from month to month for some indicators, but no month was consistently and clearly any better or worse in its predictability.

The regression analysis implied that the observed speeds followed a linear regression line to some extent, but spreading of the observations around the regression line made prediction difficult. This spreading increased as congestion progressed, thus, again, implying that prediction is harder during congestion.

Evaluating the results of the link combination is epistemically challenging. It is a novel technique, so it is hard to know whether the results are improvements or artefacts. The grand mean speed did behave as expected from the congestion levels. The new indicators should not be taken as sources of definite judgement on the functionality of the link combination procedure. It is not even known whether comparing the results with taking the mean of the indicators is a meaningful way to go about evaluating the procedure. If we assume that it is, that leaves us with the grand standard deviation, mean of PTI values and mean of SDM values. The grand standard deviation showed similar varied behaviour as it did in previous analyses of this study. The PTI might not be the best indicator to use as a guide of functionality, as argued below. This leaves the comparison between the mean of the SDM values of the 5 minute medians and the SDM of the link combination analysis. This comparison is shown in Figure 10. On three occasions there are clear differences between m(SDM)

and SDM. On most occasions the SDM is slightly higher than m(SDM). The results are similar enough that it does not weaken the functionality of the link combination procedure.

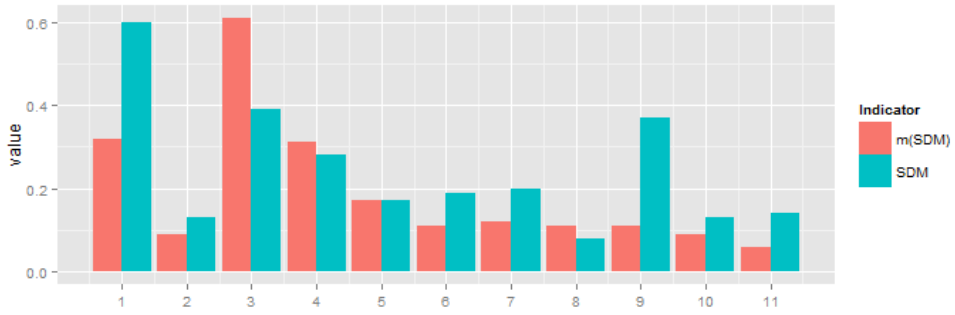

Figure 10 *Comparison of m(SDM) and SDM values from Section 3.5. Apart from comparisons 1, 3 and 9 they are very close to each other.*

Future research should specifically study the link combination procedure. This could be done with a data set where one long road segment is cut into shorter links so that we know how fast individual vehicles drove throughout the segment and on each shorter link. The link data from some set of vehicles could be combined and the results compared with actual observations of the whole segment.

One underlying issue is the definition of predictability. A driver leaving for work who knows nothing about the current state of traffic, may not be able to accurately predict the time of arrival if there is congestion on the chosen route. Here the PTI might be a good indicator as it only takes into account the current speed and free flow speed of traffic. If, on the other hand, the driver has a smartphone or navigator device that provides information on congestion along the road network, he/she might, possibly from experience, be able to take this into account when predicting the time of arrival. However, if some of the vehicles in the congestion travel much faster or slower than the mean or median speed, the driver may not be able to predict the time of arrival, because he/she may not know how fast or slow he/she will be compared to the mean or median speed. For this kind of predictability the indicators that look at the deviation of vehicle speeds, like standard deviation, MAD or the difference between fast and slow vehicles, might be better indicators than the PTI. Indicators that take into account both the mean or median speed and the deviation, such as SDM, might be seen as compromises in this regard.

All in all, the results of this study imply that the predictability of traffic worsens during the transition to congestion and during congestion itself. Exactly how much worse varies.

References

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), 764-766.

Metsäranta, H., Kiiskilä, K., Launonen, P., & Kivari, M. (2013): Matkojen ja kuljetusten palvelutaso ja tunnusluvut. Palvelutasohankkeen tuloksia vuonna 2012. Liikenneviraston tutkimuksia ja selvityksiä 4/2013. Available: http://www2.liikennevirasto.fi/julkaisut/pdf3/lts_2013-04_matkojen_ja_kuljetusten_web.pdf [Accessed 22.07.2014]

Liikennevirasto

Finnish Transport Agency

ISSN-L 1798-6656
ISSN 1798-6664
ISBN 978-952-317-053-7
www.liikennevirasto.fi